Published online 23 November 2013 *Nucleic Acids Research, 2014, Vol. 42, Database issue* ***D879–D881***
doi:10.1093/nar/gkt1164

ExoLocator—an online view into genetic makeup of vertebrate proteins

Aik Aun Khoo[1], Mario Ogrizek-Tomaš[2], Ana Bulović[2], Matija Korpar[2], Ece Gürler[1], Ivan Slijepčević[2], Mile Šikić[1,2] and Ivana Mihalek[1,*]

[1]Biomolecular Modeling and Design Division, Bioinformatics Institute 30 Biopolis Street, #07-01 Matrix, 138671 Singapore and [2]Department of Electronic Systems and Information Processing, Faculty of Electrical Engineering and Computer Science, Unska 3, 10000 Zagreb, Croatia

Received June 3, 2013; Revised October 27, 2013; Accepted October 29, 2013

ABSTRACT

ExoLocator (http://exolocator.eopsf.org) collects in a single place information needed for comparative analysis of protein-coding exons from vertebrate species. The main source of data—the genomic sequences, and the existing exon and homology annotation—is the ENSEMBL database of completed vertebrate genomes. To these, ExoLocator adds the search for ostensibly missing exons in orthologous protein pairs across species, using an extensive computational pipeline to narrow down the search region for the candidate exons and find a suitable template in the other species, as well as state-of-the-art implementations of pairwise alignment algorithms. The resulting complements of exons are organized in a way currently unique to ExoLocator: multiple sequence alignments, both on the nucleotide and on the peptide levels, clearly indicating the exon boundaries. The alignments can be inspected in the web-embedded viewer, downloaded or used on the spot to produce an estimate of conservation within orthologous sets, or functional divergence across paralogues.

INTRODUCTION

The whole-genome sequencing projects, completed (1) or still under way (2), are bringing the comparative analysis of genomic and protein sequences (3) to a whole new level of insight and reliability, giving the impetus to the field. However, assembling an exhaustive set of recognizably related sequences, be it protein or nucleotide, such that they are complete, and their source clear, remains a painstaking task. ExoLocator aims to alleviate the problem for the case for which it is currently feasible: protein coding sequences from the fully sequenced vertebrate genomes. Protein coding sequences tend to be easier to locate on the genome than the rest of the functional material therein, and the homologues from different species are also easier to faithfully align once translated to the amino acid alphabet. Working with the completed genomes enables us to establish the cognate sequences in various species that are the closest mutual homologues. These can then be used as templates for the similarity search to locate the full complement of exons for each studied gene. Finally, to organize the database in searchable chunks, we use human genome as the organizing point, the type of organization that commonly agrees with the search criteria used in biomedically motivated analysis.

The need for such data compilation available in a single place has been recognized in the community by the earlier servers of related nature (4–7), though smaller in scope than our effort presented here.

RESULTS AND THEIR PRESENTATION

Information collected in ExoLocator

ExoLocator takes ENSEMBL (1) database as its primary source of information. The data set is organized using human genome as the orientation map. All human genes annotated in the ENSEMBL as 'known' and 'protein coding' are collected, and their identifier used as a reference for the whole group of vertebrate genes annotated as orthologues (one-to-one or one-to-many) by the ENSEMBL annotation pipeline.

According to ENSEMBL, some exons do not seem to have a counterpart in closely related orthologous genes, and ExoLocator is in part an investigation into the possibility that they were overlooked in the annotation process. As an estimate of the amount of information lost, we take all canonical exons from human protein coding genes, and align them with the exons from the genes annotated as orthologous in other species. In these alignments, some 15% of expected orthologues of human exons appear absent. In our pipeline, 85% of the regions

*To whom correspondence should be addressed. Tel: +65 6478 8298; Fax: +65 6478 9048; Email: ivanam@bii.a-star.edu.sg

where we expect to find the exons (see 'Materials and Methods' section) contain explicitly indicated un-sequenced stretches of genomic sequence. Two to three percent of the missing exons are still recoverable, in full or in partial length, by search by homology.

The exons found by the pipeline are added to the overall collection, and organized in several different ways for display and downstream analysis.

ExoLocator's web interface

The database offers for download the set of protein coding exons compiled from the ENSEMBL, complemented with a straightforward homology search. It also provides the most complete reconstruction of full protein sequences we can achieve in this approach.

ExoLocator's interface provides several ways to inspect the data: as lists of exons corresponding to a gene in each species, as an alignment of orthologous proteins with the exon boundaries indicated, as an alignment of within-species paralogues, or as an alignment of alternative splices. The last option is available only for the cases that have Consensus CDS annotation (8). The alignments are available at the nucleotide and amino acid levels. The visualization of the alignment is provided by the browser embedded JalView alignment viewer (9). The orthologue alignment comes with a set of notes, detailing the list of the exons it contains—their position in the gene, the source: the ENSEMBL itself, Havana annotation project (10) or the similarity search using the closest detectable homologue (see 'Materials and Methods' section).

The search in ExoLocator can be done by providing the ENSEMBL identifier, pasting in the sequence on the protein level or through a limited name resolution search.

MATERIALS AND METHODS

The original exon set available at ENSEMBL, release e73, our main source of raw genomic data, was assembled through a combination of de novo gene detection and heuristic search (BLAST) by similarity. To that arsenal of methods we have added a pairwise alignment (or similarity search) algorithm by Edgar (11), and our in-house implementation of a hardware accelerated version of Smith–Waterman search (12).

In addition to being an extensive exercise in mining the ENSMBL core data, the database also provides an insight into the extent to which the number of known exons can be extended by optimal sequence alignment (applied to detection of homologous sequences across species). To establish the search pipeline, implemented in Python 2.6, we had to make several decisions, and develop appropriate software.

Pipeline description

The first decision we make is to select a canonical set of exons for each human gene in ENSEMBL to use as the reference points in our search. Where exons overlap or disagree, exons annotated as 'known' with the greatest length and coverage are chosen over the others. This we do by modeling exons as nodes in a directed acyclic graph, with edges going from overlapping exons with greater quality—measured by the strength of the annotation (Havana over ENSEMBL; strongly supported splice signal over none), the length and the similarity to a known template to existing species—to lesser, then taking the set of nodes with no incoming edges as our model set of exons for the gene.

Then, to each human exon we attach a map to 'master' exons in the other species from the corresponding genes in other species. The maps are further reconciled in a full-length protein alignment, to detect and accommodate the cases of different intron positioning across species. An *ad hoc* pairwise aligner that respects exon boundaries is used for the purpose. For the final alignment on the multiple sequence level we use MAFFT alignment utility (13).

Search for missing exons

To detect a missing exon we align a target vertebrate set of exons corresponding to a single gene to the most convincingly homologous set in human. To relate exons to their parent gene we again rely on the annotation provided by the ENSEMBL. The alignment provides the boundaries on the target gene for the search for the missing exon. Next, we need to choose a template from the species that has the exon annotated, and is in some sense the nearest to the species with the exon annotation missing. For that purpose we use the taxonomy tree available at the NCBI's Taxonomy Web site (14). We traverse the tree to look for the taxonomically closest species that has an exon mapping to the human at the expected place, and use it as a template for the sequence similarity search.

Finally, to detect the region of homology, we use an advanced CPU implementation of a heuristic search, and GPU enabled Smith–Waterman algorithm. The implementations of the latter available in the public domain are not capable of handling the sizes for the input sequences we have at hand, and therefore we use our own implementation in which the problem is divided into smaller chunks in a way amenable to graphics card acceleration (https://github.com/mkorpar/swSharp). The exons that ExoLocator reports satisfy two criteria: their translation must be longer then three residues, and similarity by a Tanimoto-like similarity measure $\sqrt{(S_{12}^2/L_1L_2)}$, where L_1 and L_2 are the lengths of the template and the candidate exon, and S_{12} is the similarity weighted length of the common aligned positions, must be larger than 1/3.

Known problems and caveats

Some otherwise interesting peculiarities of animal genomes complicate systematic analysis of the sort we undertook here. Thus we do not attempt to resolve the cases of overlapping genes on the same strand, of which we detected several hundred cases (some of these, though, might be duplicate entries in the source database we are using). As we rely on the ENSEMBL pipeline for the annotation of at least approximate exon location, when the whole gene is unannotated in the species, it will be missing in our search too. Also, the detection

by similarity that we use does not allow us to decide on the precise location of the gene boundary. We use MaxEntScan (15), a lightweight implementation of a strong statistical tool to try to estimate the likelihood that the predicted exon has a proper splice signal. MaxEntScan in its currently available parametrization works the best for mammalian sequences of introns being spliced out by the major spliceosome. In these cases we use MaxEntScan to decide on the boundary and the possible phase of the exon, and provide the score MaxEntScan assigns in the accompanying notes.

Database implementation

The database that is accessible via Internet is a relatively straightforward MySQL application with a small number of tables storing the information about the exon coordinates, sequences and the source of the annotation. The processing pipeline used to fill the database, however, is a much more complex Python implementation sourcing the data from the local versions of the core ENSEMBL databases. The Web interface for the database was implemented in Play! Framework (http://www.playframework.org). With the full cycle of data processing being rather time-consuming, the database will be on a semiannual update schedule.

CONCLUSION AND OUTLOOK

At its current stage, ExoLocator aims to balance the goal of giving the complete picture of possible coding exons, with the need to stay grounded in terms of the verifiability of the actual function of the sequences it collects. Thus it relies on the ENSEMBL's annotation of 'known' ('known' here being the actual annotation term used, hence the quotes) human exons as the anchor for the search and for the results presentation. In certain cases, it seems that the data from species other than human argue for a different annotation, but we deliberately choose to stay away from any reinterpretation of the existing data. Rather, we envision the database to function as a shortcut to quick retrieval of the established sequences of well-documented human exons and their closest counterparts in the other species, complemented by the putative exon set collected by a reliable search utility. The hope is that, rather than as an interpretative tool, ExoLocator will be understood and used as a resource, ultimately leading to fuller understanding of the complex mechanism of gene function, alternative splicing and translation.

FUNDING

Biomedical Research Council of Agency for Science, Technology and Research Singapore. Funding for open access charge: Agency for Science, Technology and Research, Singapore.

Conflict of interest statement. None declared.

REFERENCES

1. Flicek,P., Amode,M., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) ENSEMBL 2012. *Nucleic Acids Res.*, **40**, D84–D90.
2. Haussler,D., O'Brien,S.J., Ryder,O.A., Barker,F.K., Clamp,M., Crawford,A.J., Hanner,R., Hanotte,O., Johnson,W.E., McGuire,J.A. *et al.* (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10000 vertebrate species. *J. Hered.*, **100**, 659–674.
3. Liberles,D., Teichmann,S., Bahar,I., Bastolla,U., Bloom,J., Bornberg-Bauer,E., Colwell,L., de Koning,A., Dokholyan,N., Echave,J. *et al.* (2012) The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci.*, **21**, 769785.
4. Saxonov,S., Daizadeh,I., Fedorov,A. and Gilbert,W. (2000) EID: the Exon–Intron Databasean exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.*, **28**, 185–190.
5. Alekseyenko,A., Kim,N. and Lee,C. (2007) Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. *RNA*, **13**, 661–670.
6. Mollet,I., Ben-Dov,C., Felício-Silva,D., Grosso,A., Eleutério,P., Alves,R., Staller,R., Silva,T. and Carmo-Fonseca,M. (2010) Unconstrained mining of transcript data reveals increased alternative splicing complexity in the human transcriptome. *Nucleic Acids Res.*, **38**, 4740–4754.
7. Busch,A. and Hertel,K.J. (2013) HEXEvent: a database of human EXon splicing Events. *Nucleic Acids Res.*, **41**, D118–D124.
8. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E., Ruef,B.J. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.
9. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview Version 2a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
10. Wilming,L., Gilbert,J., Howe,K., Trevanion,S., Hubbard,T. and Harrow,J. (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.*, **36(Suppl. 1)**, D753–D760.
11. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
12. Korpar,M. and Sikic,M. (2013) SW#–GPU-enabled exact alignments on genome scale. *Bioinformatics*, **29**, 2494–2495.
13. Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
14. Federhen,S. (2012) The NCBI taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
15. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.